

Gaussian Rate-Distortion via Sparse Linear Regression over Compact Dictionaries

Ramji Venkataramanan
Dept. of Electrical Engineering
Yale University, USA
Email: ramji.venkataramanan@yale.edu

Antony Joseph
Dept. of Statistics
Yale University, USA
Email: antony.joseph@yale.edu

Sekhar Tatikonda
Dept. of Electrical Engineering
Yale University, USA
Email: sekhar.tatikonda@yale.edu

Abstract—We study a class of codes for compressing memoryless Gaussian sources, designed using the statistical framework of high-dimensional linear regression. Codewords are linear combinations of subsets of columns of a design matrix. With maximum-likelihood encoding we show that such a codebook can attain the rate-distortion function with the optimal error-exponent, for all distortions below a specified value. The structure of the codebook is motivated by an analogous construction proposed recently by Barron and Joseph for communication over an AWGN channel.

I. INTRODUCTION

One of the important outstanding problems in information theory is the development of practical codes for lossy compression of general sources at rates approaching Shannon's rate-distortion bound. In this paper, we study the compression of memoryless Gaussian sources using a class of codes constructed based on the statistical framework of high-dimensional linear regression. The codebook consists of codewords that are sparse linear combinations of columns of an $n \times N$ design matrix or 'dictionary', where n is the block-length and N is a low-order polynomial in n . Dubbed Sparse Superposition Codes or Sparse Regression Codes (SPARC), these codes are motivated by an analogous construction proposed recently by Barron and Joseph for communication over an AWGN channel [1], [2]. The structure of the codebook enables the design of computationally efficient encoders based on the rich theory on sparse linear regression. Here, we study the performance of these codes under maximum-likelihood (ML) encoding. The design of feasible encoders will be discussed in future work.

Sparse regression codes for compressing Gaussian sources were first considered in [3] where some preliminary results were presented. In this paper, we analyze the performance of these codes under ML encoding and show that they can achieve the distortion-rate bound with the optimal error exponent for all rates above a specified value (approximately 1.15 bits/sample). The proof uses Suen's inequality [4], a bound on the tail probability of a sum of dependent indicator random variables. This technique may be of independent interest and useful in other problems in information theory.

We lay down some notation before proceeding further. Upper-case letters are used to denote random variables, lower-case for their realizations, and bold-face letters to denote

random vectors and matrices. All vectors have length n . The source sequence is denoted by $\mathbf{S} \triangleq (S_1, \dots, S_n)$, and the reconstruction sequence by $\hat{\mathbf{S}} \triangleq (\hat{S}_1, \dots, \hat{S}_n)$. $\|\mathbf{X}\|$ denotes the l^2 -norm of vector \mathbf{X} , and $|\mathbf{X}| = \|\mathbf{X}\|/\sqrt{n}$ is the normalized version. We use natural logarithms, so entropy is measured in nats. $f(x) = o(g(x))$ means $\lim_{x \rightarrow \infty} f(x)/g(x) = 0$; $f(x) = \Theta(g(x))$ means $f(x)/g(x)$ asymptotically lies in an interval $[k_1, k_2]$ for some constants $k_1, k_2 > 0$.

Consider an i.i.d Gaussian source S with mean 0 and variance σ^2 . A rate-distortion codebook with rate R and block length n is a set of e^{nR} length- n codewords, denoted $\{\hat{\mathbf{S}}(1), \dots, \hat{\mathbf{S}}(e^{nR})\}$. The quality of reconstruction is measured through a mean-squared error distortion criterion

$$d_n(\mathbf{S}, \hat{\mathbf{S}}) = \|\mathbf{S} - \hat{\mathbf{S}}\|^2 = \frac{1}{n} \sum_{i=1}^n (S_i - \hat{S}_i)^2,$$

where $\hat{\mathbf{S}}$ is the codeword chosen to represent the source sequence \mathbf{S} . For this distortion criterion, an optimal (maximum-likelihood) encoder maps each source sequence to the codeword nearest to it in Euclidean distance. The rate-distortion function $R^*(D)$, the minimum rate for which the distortion can be bounded by D with high-probability, is given by [5]

$$R^*(D) = \min_{p_{\mathbf{S}|\mathbf{S}}: E(\mathbf{S} - \hat{\mathbf{S}})^2 \leq D} I(\mathbf{S}; \hat{\mathbf{S}}) = \frac{1}{2} \log \frac{\sigma^2}{D} \text{ nats/sample.} \quad (1)$$

This rate can be achieved through Shannon-style random codebook selection: pick each codeword independently as an i.i.d Gaussian random vector distributed as $\text{Normal}(0, \sigma^2 - D)$.

Lattice-based codes for Gaussian vector quantization have been widely studied, e.g [6], [7]. There are computationally efficient quantizers for certain classes of lattice codes, but the high-dimensional lattices needed to approach the rate-distortion bound have exponential encoding complexity [7]. We also note that for sources with finite alphabet, various coding techniques have been proposed recently to approach the rate-distortion bound with computationally feasible encoding and decoding [8]–[11].

II. SPARSE REGRESSION CODES

A sparse regression code (SPARC) is defined in terms of a design matrix \mathbf{A} of dimension $n \times ML$ whose entries are zero mean i.i.d Gaussian random variables with variance $\frac{\sigma^2 - D}{L}$,

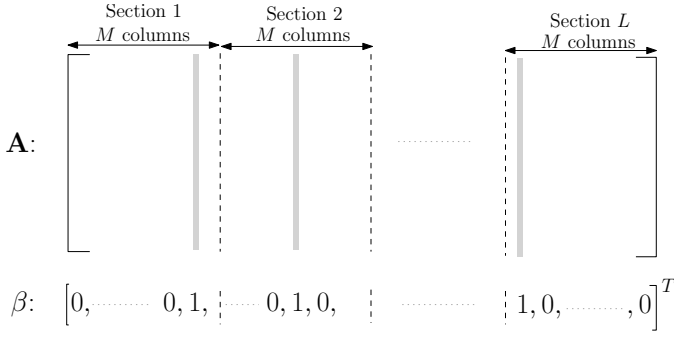


Fig. 1: \mathbf{A} is an $n \times ML$ matrix and β is a $ML \times 1$ binary vector. The positions of the ones in β correspond to the gray columns of \mathbf{A} which add to form the codeword $\mathbf{A}\beta$.

where the constant a^2 will be specified in the sequel. Here n is the block length and M and L are integers whose values will be specified shortly in terms of n and the rate R . As shown in Figure 1, one can think of the matrix \mathbf{A} as composed of L sections with M columns each. Each codeword is the sum of L columns, with one column from each section. More formally, a codeword can be expressed as $\mathbf{A}\beta$, where β is a binary-valued $ML \times 1$ vector $(\beta_1, \dots, \beta_{ML})$ with the following property: there is exactly one non-zero β_i for $1 \leq i \leq M$, one non-zero β_i for $M+1 \leq i \leq 2M$, and so forth. Denote the set of all β 's that satisfy this property by $\mathcal{B}_{M,L}$.

Maximum-likelihood Encoder: This is defined by a mapping $g : \mathbb{R}^n \rightarrow \mathcal{B}_{M,L}$. Given the source sequence \mathbf{S}^n , the encoder determines the β that produces the codeword closest in Euclidean distance, i.e.,

$$g(\mathbf{S}) = \underset{\beta \in \mathcal{B}_{M,L}}{\operatorname{argmin}} \|\mathbf{S} - \mathbf{A}\beta\|.$$

Decoder: This is a mapping $h : \mathcal{B}_{M,L} \rightarrow \mathbb{R}^n$. On receiving $\beta \in \mathcal{B}_{M,L}$ from the encoder, the decoder produces reconstruction $h(\beta) = \mathbf{A}\beta$.

Since there are M columns in each of the L sections, the total number of codewords is M^L . To obtain a compression rate of R nats/sample, we therefore need

$$M^L = e^{nR}. \quad (2)$$

There are several choices for the pair (M, L) which satisfy this. For example, $L = 1$ and $M = e^{nR}$ recovers the Shannon-style random codebook in which the number of columns in the dictionary \mathbf{A} is e^{nR} , i.e., exponential in n . For our constructions, we choose $M = L^b$ for some $b > 1$ so that (2) implies

$$L \log L = nR/b. \quad (3)$$

Thus L is now $\Theta\left(\frac{n}{\log n}\right)$, and the number of columns ML in the dictionary \mathbf{A} is now $\Theta\left(\frac{n}{\log n}\right)^{b+1}$, a polynomial in n . This reduction in dictionary complexity can be harnessed to develop computationally efficient encoders for the sparse regression code. We note that the code structure automatically yields low decoding complexity.

Since each codeword in a SPARC is a sum of L columns of \mathbf{A} (one from each section), codewords sharing one or more common columns in the sum will be dependent. Also, SPARCs are not linear codes since the sum of two codewords does not equal another codeword in general.

III. MAIN RESULT

We begin with some background on error exponents.

The probability of error at distortion-level D of a rate-distortion code \mathcal{C}_n with block length n and encoder and decoder mappings g, h is

$$P_e(\mathcal{C}_n, D) = P(|\mathbf{S} - h(g(\mathbf{S}))|^2 > D).$$

Definition 1: The error exponent at distortion-level D of a sequence of rate R codes $\{\mathcal{C}_n\}_{n=1,2,\dots}$ is given by

$$r(R, D) = -\limsup_{n \rightarrow \infty} \frac{1}{n} \log P_e(\mathcal{C}_n, D). \quad (4)$$

The optimal error exponent for a rate-distortion pair (R, D) is the supremum of the error exponents over all sequences of codes with rate R , at distortion-level D .

The error-exponent describes the asymptotic behavior of the probability of error; bounds on the probability of error for finite block lengths were obtained in [12], [13]. The optimal error exponent was obtained by Marton [14] for discrete memoryless sources and was extended to Gaussian sources by Ihara and Kubo [15].

Fact 1: [15] For an i.i.d Gaussian source distributed as $\text{Normal}(0, \sigma^2)$ and mean-squared error distortion criterion, the optimal error exponent at rate R and distortion-level D is

$$r^*(D, R) = \begin{cases} \frac{1}{2} \left(\frac{\rho^2}{\sigma^2} - 1 - \log \frac{\rho^2}{\sigma^2} \right) & R > R^*(D) \\ 0 & R \leq R^*(D) \end{cases} \quad (5)$$

where ρ^2 is determined by

$$R = \frac{1}{2} \log \frac{\rho^2}{D}. \quad (6)$$

For $R > R^*(D)$, the exponent in (5) is the Kullback-Leibler divergence between two zero-mean Gaussian distributions, the first with variance ρ^2 and the second with variance σ^2 . [15] shows that at rate R , we can compress all sequences which have empirical variance less than ρ^2 to within distortion D with double-exponentially decaying probability of error. Consequently, the dominant error event is obtaining a source sequence with empirical variance greater than ρ^2 , which has exponent given by (5).

The main result of our paper is the following.

Theorem 1: Fix a rate R and target distortion D such that $\sigma^2/D > x^*$, where $x^* \approx 4.913$ is the solution of the equation

$$\frac{1}{2} \log x = \left(1 - \frac{1}{x}\right).$$

Fix $b > \frac{4R}{R - (1 - D/\rho^2)}$, where ρ^2 is determined by (6). For every positive integer n , let $M_n = L_n^b$ where L_n is determined by (3). Then there exists a sequence $\mathcal{C} = \{\mathcal{C}_n\}_{n=1,2,\dots}$ of rate

R sparse regression codes - with code \mathcal{C}_n defined by an $n \times M_n L_n$ design matrix - that attains the optimal error exponent for distortion-level D given by (5).

Remark: The minimum value of b specified by the theorem enables us to construct SPARCs with the *optimal* error exponent. The proof also shows that we can construct SPARCs which achieve the rate-distortion function for $b > \frac{3R}{R-(1-D/\rho^2)}$, with probability of error that decays sub-exponentially in n when b is less than $4R/(R-(1-D/\rho^2))$.

IV. PROOF OF THEOREM 1

Due to space constraints, we omit some details in the proof which will be included in a longer version of this paper. Given rate $R > R^*(D)$, let ρ^2 be determined by (6). For each $a^2 < \rho^2$, we will show that there exists a family of SPARCs that achieves the error exponent $\frac{1}{2} \left(\frac{a^2}{\sigma^2} - 1 - \log \frac{a^2}{\sigma^2} \right)$, thereby proving the theorem.

Code Construction: For each block length n , pick L as specified by (3) and $M = L^b$. Construct an $n \times ML$ design matrix \mathbf{A} with entries drawn i.i.d Normal($0, \frac{a^2-D}{L}$). The codebook consists of all the vectors $\mathbf{A}\beta$, where $\beta \in \mathcal{B}_{M,L}$.

Encoding and Decoding: If the source sequence \mathbf{S} is such that $|\mathbf{S}|^2 \geq a^2$, then the encoder declares error. Else, it finds

$$\hat{\beta} \triangleq g(\mathbf{S}) = \operatorname{argmin}_{\beta \in \mathcal{B}_{M,L}} \|\mathbf{S} - \mathbf{A}\beta\|.$$

The decoder receives $\hat{\beta}$ and reconstructs $\hat{\mathbf{S}} = \mathbf{A}\hat{\beta}$.

Error Analysis: Denoting the probability of error for this random code by $P_{e,n}$, we have

$$\begin{aligned} P_{e,n} &\leq 1 \cdot P(|\mathbf{S}|^2 \geq a^2) + \int_0^{a^2} P(\mathcal{E}(\mathbf{S}) \mid |\mathbf{S}|^2 = z^2) d\nu(z^2) \\ &\leq P(|\mathbf{S}|^2 \geq a^2) + \max_{z^2 \in (0, a^2)} P(\mathcal{E}(\mathbf{S}) \mid |\mathbf{S}|^2 = z^2). \end{aligned} \quad (7)$$

where $\mathcal{E}(\mathbf{S})$ is the event that the minimum of $|\mathbf{S} - \mathbf{A}\beta|^2$ over $\beta \in \mathcal{B}_{M,L}$ is greater than D , and $\nu(|\mathbf{S}|^2)$ is the distribution of the random variable $|\mathbf{S}|^2$. The asymptotic behavior of the first term above is straightforward to analyze and is given by the following lemma, obtained through a direct application of Cramér's large-deviation theorem [16].

Lemma 1:

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log P(|\mathbf{S}|^2 > a^2) = \frac{1}{2} \left(\frac{a^2}{\sigma^2} - 1 - \log \frac{a^2}{\sigma^2} \right).$$

The rest of the proof is devoted to bounding the second term in (7). Recall that

$$\begin{aligned} P(\mathcal{E}(\mathbf{S}) \mid |\mathbf{S}|^2 = z^2) \\ = P(|\hat{\mathbf{S}}(i) - \mathbf{S}|^2 \geq D, i = 1, \dots, e^{nR} \mid |\mathbf{S}|^2 = z^2) \end{aligned} \quad (8)$$

where $\hat{\mathbf{S}}(i)$ is the i th codeword in the sparse regression codebook. We now define indicator random variables $U_i(\mathbf{S})$ for $i = 1, \dots, e^{nR}$ as follows:

$$U_i(\mathbf{S}) = \begin{cases} 1 & \text{if } |\hat{\mathbf{S}}(i) - \mathbf{S}|^2 < D, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

From (8), it is seen that

$$P(\mathcal{E}(\mathbf{S}) \mid |\mathbf{S}|^2 = z^2) = P \left(\sum_{i=1}^{e^{nR}} U_i(\mathbf{S}) = 0 \mid |\mathbf{S}|^2 = z^2 \right). \quad (10)$$

For a fixed \mathbf{S} , the $U_i(\mathbf{S})$'s are dependent. Suppose that the codewords $\hat{\mathbf{S}}(i), \hat{\mathbf{S}}(j)$ respectively correspond to the binary vectors $\hat{\beta}(i), \hat{\beta}(j) \in \mathcal{B}_{M,L}$. Recall that each vector in $\mathcal{B}_{M,L}$ is uniquely defined by the position of the 1 in each of the L sections. If $\hat{\beta}(i)$ and $\hat{\beta}(j)$ overlap in r of their '1 positions', then the column sums forming codewords $\hat{\mathbf{S}}(i)$ and $\hat{\mathbf{S}}(j)$ will share r common terms.

For each codeword $\hat{\mathbf{S}}(i)$, there are $\binom{L}{r}(M-1)^{L-r}$ other codewords which share exactly r common terms with $\hat{\mathbf{S}}(i)$, for $0 \leq r \leq L-1$. In particular, there are $(M-1)^L$ codewords that are pairwise independent of $\hat{\mathbf{S}}(i)$. We now obtain an upper bound for the probability in (10) using Suen's correlation inequality [4]. First, some definitions.

Definition 2 (Dependency Graphs [4]): Let $\{U_i\}_{i \in \mathcal{I}}$ be a family of random variables (defined on a common probability space). A dependency graph for $\{U_i\}$ is any graph Γ with vertex set $V(\Gamma) = \mathcal{I}$ whose set of edges satisfies the following property: if A and B are two disjoint subsets of \mathcal{I} such that there are no edges with one vertex in A and the other in B , then the families $\{U_i\}_{i \in A}$ and $\{U_i\}_{i \in B}$ are independent.

Fact 2: [4, Example 1.5, p.11] Suppose $\{Y_\alpha\}_{\alpha \in \mathcal{A}}$ is a family of independent random variables, and each $U_i, i \in \mathcal{I}$ is a function of the variables $\{Y_\alpha\}_{\alpha \in A_i}$ for some subset $A_i \subseteq \mathcal{A}$. Then the graph with vertex set \mathcal{I} and edge set $\{ij : A_i \cap A_j \neq \emptyset\}$ is a dependency graph for $\{U_i\}_{i \in \mathcal{I}}$.

Remark 1: The graph Γ with vertex set $V(\Gamma) = \{1, \dots, e^{nR}\}$ and edge set $e(\Gamma)$ given by

$$\{ij : i \neq j \text{ and } \hat{\mathbf{S}}(i), \hat{\mathbf{S}}(j) \text{ share at least one common term}\}$$

is a dependency graph for the family $\{U_i(\mathbf{S})\}_{i=1}^{e^{nR}}$, for each fixed \mathbf{S} . This follows from Fact 2 by recognizing that each U_i is a function of a subset of the columns of the matrix \mathbf{A} and the columns of \mathbf{A} are picked independently in the code construction.

Fact 3 (Suen's Inequality [4]): Let $U_i \sim \text{Bern}(p_i), i \in \mathcal{I}$, be a finite family of Bernoulli random variables having a dependency graph Γ . Write $i \sim j$ if ij is an edge in Γ . Define

$$\lambda = \sum_{i \in \mathcal{I}} \mathbb{E}U_i, \quad \Delta = \frac{1}{2} \sum_{i \in \mathcal{I}} \sum_{j \sim i} \mathbb{E}(U_i U_j), \quad \delta = \max_{i \in \mathcal{I}} \sum_{k \sim i} \mathbb{E}U_k.$$

Then

$$P \left(\sum_{i \in \mathcal{I}} U_i = 0 \right) \leq \exp \left(- \min \left\{ \frac{\lambda}{2}, \frac{\lambda}{6\delta}, \frac{\lambda^2}{8\Delta} \right\} \right).$$

We now apply this inequality with the dependency graph specified in Remark 1 to compute an upper bound for (10).

First term $\lambda/2$: Since each codeword is the sum of L columns of \mathbf{A} whose entries are i.i.d Normal($0, a^2 - D$),

$\mathbb{E}(U_i(\mathbf{S}))$ does not depend on i . For any fixed \mathbf{S} with $|\mathbf{S}|^2 = z^2$, we have

$$\lambda = \sum_{i=1}^{e^{nR}} \mathbb{E}(U_i(\mathbf{S})) = e^{nR} P(U_1(\mathbf{S}) = 1 \mid |\mathbf{S}|^2 = z^2). \quad (11)$$

Using the strong version of Cramér's large-deviation theorem by Bahadur and Rao [17], we can obtain the following lemma.

Lemma 2: For all sufficiently large n and $z^2 \in (0, a^2)$,

$$\begin{aligned} P(U_1(\mathbf{S}) = 1 \mid |\mathbf{S}|^2 = z^2) &\geq P(U_1(\mathbf{S}) = 1 \mid |\mathbf{S}|^2 = a^2) \\ &\geq \frac{1}{\kappa\sqrt{n}} e^{-\frac{n}{2} \log \frac{a^2}{D}} = e^{-n \left(\frac{1}{2} \log \frac{a^2}{D} + \frac{\log n}{2n} + \frac{\kappa}{n} \right)} \end{aligned}$$

for some constant $\kappa > 0$.

We thus have a lower bound on λ for sufficiently large n .

Second term λ/δ : Due to the symmetry of the code construction,

$$\begin{aligned} \delta &\triangleq \max_{i \in \{1, \dots, e^{nR}\}} \sum_{k \sim i} P(U_k(\mathbf{S}) = 1 \mid |\mathbf{S}|^2 = z^2) \\ &= \sum_{k \sim i} P(U_k(\mathbf{S}) = 1 \mid |\mathbf{S}|^2 = z^2) \quad \forall i \in \{1, \dots, e^{nR}\} \\ &= \sum_{r=1}^{L-1} \binom{L}{r} (M-1)^{L-r} \cdot P(U_1(\mathbf{S}) = 1 \mid |\mathbf{S}|^2 = z^2) \\ &= (M^L - 1 - (M-1)^L) P(U_1(\mathbf{S}) = 1 \mid |\mathbf{S}|^2 = z^2). \end{aligned} \quad (12)$$

Using this together with the expression for λ in (11), we have

$$\frac{\lambda}{\delta} = \frac{M^L}{M^L - 1 - (M-1)^L} = \frac{1}{1 - L^{-bL} - [(1 - L^{-b})^{L^b}]^{L^{1-b}}} \quad (13)$$

where we have used $M = L^b$. Since $(1 - L^{-b})^{L^b} \rightarrow e^{-1}$, we can show using a Taylor expansion that for L sufficiently large

$$\frac{\lambda}{\delta} \geq \frac{1}{L^{-(b-1)} - \frac{1}{4}L^{-2(b-1)} + o(L^{-2(b-1)})} \geq L^{b-1}. \quad (14)$$

Third Term λ^2/Δ : We lower bound λ^2/Δ by obtaining a lower bound for λ^2 using Lemma 2, and an upper bound for the denominator Δ as follows.¹

$$\begin{aligned} \Delta &= \frac{1}{2} \sum_{i=1}^{e^{nR}} \sum_{j \sim i} \mathbb{E}(U_i(\mathbf{S})U_j(\mathbf{S}) \mid |\mathbf{S}|^2 = a^2) = \\ &= \frac{e^{nR}}{2} \sum_{r=1}^{L-1} \binom{L}{r} (M-1)^{L-r} P(U_i(\mathbf{S}) = U_j(\mathbf{S}) = 1 \mid E_{ij}(r), |\mathbf{S}|^2 = a^2) \end{aligned} \quad (15)$$

¹Here we directly lower bound on λ^2/Δ for $|\mathbf{S}|^2 = a^2$. Formally, a lower bound on λ^2/Δ can be obtained using similar steps for $|\mathbf{S}|^2 = z^2$ for $z^2 \in (0, a^2)$, and it can be shown to be decreasing in z^2 .

where $E_{ij}(r)$ is the event that $\hat{\mathbf{S}}(i), \hat{\mathbf{S}}(j)$ have exactly r common terms. We have

$$\begin{aligned} P(U_i(\mathbf{S}) = 1, U_j(\mathbf{S}) = 1 \mid E_{ij}(r), |\mathbf{S}|^2 = a^2) \\ = P(|\hat{\mathbf{S}}(i) - \mathbf{S}|^2 \leq D, |\hat{\mathbf{S}}(j) - \mathbf{S}|^2 \leq D \mid E_{ij}(r), |\mathbf{S}|^2 = a^2) \\ = P\left(\frac{1}{n} \sum_{k=1}^n (\hat{S}_k(i) - a)^2 \leq D, \frac{1}{n} \sum_{k=1}^n (\hat{S}_k(j) - a)^2 \leq D\right), \end{aligned} \quad (16)$$

where the third equality is due to the fact that $(\hat{\mathbf{S}}(i), \hat{\mathbf{S}}(j))$ has the same joint distribution as $(\mathbf{O}\hat{\mathbf{S}}(i), \mathbf{O}\hat{\mathbf{S}}(j))$ for any orthogonal (rotation) matrix \mathbf{O} . The $(\hat{S}_k(i), \hat{S}_k(j))$ pairs are i.i.d across k , and each is jointly Gaussian with zero-mean vector and covariance matrix

$$K_r = (a^2 - D) \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}, \quad \text{where } \alpha = \frac{r}{L} \quad (17)$$

when $\hat{\mathbf{S}}(i), \hat{\mathbf{S}}(j)$ share r common terms. Using a two-dimensional Chernoff bound, we have $\forall u, t < 0$

$$\begin{aligned} \frac{1}{n} \log P\left(\sum_{k=1}^n \frac{(\hat{S}_k(i) - a)^2}{n} \leq D, \sum_{k=1}^n \frac{(\hat{S}_k(j) - a)^2}{n} \leq D \mid E_{ij}(r)\right) \\ \leq \log \mathbb{E}\left(e^{u(\hat{S}(i)-a)^2 + t(\hat{S}(j)-a)^2}\right) - (u+t)D = -C_\alpha(u, t) \end{aligned} \quad (18)$$

where

$$\begin{aligned} C_\alpha(u, t) &= (u+t)D - \frac{a^2(u+t - 2\gamma^2 ut(1-\alpha^2))}{1 - 2\gamma^2 u(1-\alpha^2)} \\ &\quad - \frac{2a^2\gamma^2(u\alpha + t - 2\gamma^2 ut(1-\alpha^2))^2}{(1 - 2\gamma^2 u(1-\alpha^2))(1 - 2\gamma^2(u+t) + 4\gamma^4 ut(1-\alpha^2))} \\ &\quad + \frac{1}{2} \log(1 - 2\gamma^2(u+t) + 4\gamma^4 ut(1-\alpha^2)) \end{aligned} \quad (19)$$

with $\gamma^2 \triangleq a^2 - D$.

Using (18) in (16) and then in (15), and using Lemma 2 to bound λ^2 , we obtain

$$\frac{\lambda^2}{\Delta} > \frac{\exp\left[2n(R - \frac{1}{2} \log \frac{a^2}{D} - \gamma_n)\right]}{e^{nR} \sum_{r=1}^{L-1} \binom{L}{r} (M-1)^{L-r} e^{-nC_\alpha(u, t)}} \quad (20)$$

where

$$\gamma_n = \log n/2n + \kappa/n. \quad (21)$$

In the sequel, we will use κ to denote a generic positive constant whose exact value is not needed. Using \mathcal{A}_L to denote the set $\{1/L, 2/L, \dots, (L-1)/L\}$, α to denote $\frac{r}{L}$, and noting that $M^L = e^{nR}$, we have

$$\begin{aligned} \frac{\lambda^2}{\Delta} &> \frac{M^L \exp[-n(\log(a^2/D) + 2\gamma_n)]}{\sum_{\alpha \in \mathcal{A}_L} \binom{L}{L\alpha} (M-1)^{L(1-\alpha)} e^{-nC_\alpha(u, t)}} \\ &> \frac{\exp[-n(\log(a^2/D) + 2\gamma_n)]}{(L-1) \max_{\alpha \in \mathcal{A}_L} \binom{L}{L\alpha} M^{-\alpha L} e^{-nC_\alpha(u, t)}}. \end{aligned} \quad (22)$$

Substituting $M = L^b$ and taking logarithms, we get

$$\log \frac{\lambda^2}{\Delta} > \min_{\alpha \in \mathcal{A}_L} \left\{ (bL\alpha - 1) \log L - \log \binom{L}{L\alpha} - n(\log(a^2/D) + 2\gamma_n - C_\alpha(u, t)) \right\}. \quad (23)$$

Dividing through by $L \log L$ and using the relation (3) as well as the definition (21) for γ_n , we get

$$\frac{\log(\lambda^2/\Delta)}{L \log L} > \min_{\alpha \in \mathcal{A}_L} \left\{ b\alpha + \frac{b}{R}(C_\alpha(u, t) - \log(a^2/D)) - \frac{1}{L} \right. \\ \left. - \frac{\log(L \log L)}{L \log L} - \frac{\log(L_\alpha)}{L \log L} - \frac{\kappa}{L \log L} \right\}. \quad (24)$$

We need the right side of the above to be positive since we want λ^2/Δ to grow with L . For this, we need:

$$b > \frac{\frac{2}{L} + \frac{\log(L_\alpha)}{L \log L} + \frac{\kappa + \log \log L}{L \log L}}{\alpha + \frac{C_\alpha(u, t) - \log(a^2/D)}{R}}, \quad \forall \alpha \in \mathcal{A}_L. \quad (25)$$

Further, we need the denominator of (25) to be positive. Using $u = t = -\frac{1}{2D(1+\alpha)}$ for $C_\alpha(u, t)$, we can show that (25) implies the following simplified condition for sufficiently large L :

$$b > b_{\min} \triangleq \frac{3R}{R - (1 - D/a^2)}. \quad (26)$$

When $b > b_{\min}$, the right side of (24) will be strictly positive for large enough L . Since a^2 is any number less than ρ^2 where $R = \frac{1}{2} \log \frac{\rho^2}{D}$, the condition for the denominator to be positive is

$$\frac{1}{2} \log \frac{a^2}{D} > 1 - \frac{D}{a^2}. \quad (27)$$

This is satisfied whenever $a^2/D > x^*$ as required by the theorem. Thus for large enough L , (24) becomes

$$\frac{\log(\lambda^2/\Delta)}{L \log L} > \frac{1}{L}(b - b_{\min}) \left(1 - \frac{1}{R}(1 - \frac{D}{a^2})\right). \quad (28)$$

Therefore for sufficiently large L ,

$$\frac{\lambda^2}{\Delta} > L^{(b-b_{\min})(1-\frac{1}{R}(1-\frac{D}{a^2}))}. \quad (29)$$

Combining the bounds obtained above for each of the three terms, we have for sufficiently large n ,

$$P\left(\sum_{i=1}^{e^{nR}} U_i(\mathbf{S}) = 0\right) \leq e^{-n \min\{T_1, T_2, T_3\}} \quad (30)$$

where

$$T_1 > e^{n(R - \frac{1}{2} \log \frac{a^2}{D} - \frac{\log n}{2n} - \frac{\kappa}{n})}, \quad T_2 > L^{b-1}, \quad (31) \\ T_3 > L^{(b-b_{\min})(1-\frac{1}{R}(1-\frac{D}{a^2}))}.$$

Using this in (7), we obtain

$$P_{e,n} \leq P(|\mathbf{S}|^2 \geq a^2) + \max_{z^2 \in (0, a^2)} P(\mathcal{E}(\mathbf{S}) \mid |\mathbf{S}|^2 = z^2) \\ < e^{-nT_0} + e^{-n \min\{T_1, T_2, T_3\}} \quad (32)$$

where $T_0 = \frac{1}{2}(\frac{a^2}{\sigma^2} - 1 - \log \frac{a^2}{\sigma^2})$ from Lemma 1. Since $R = \frac{1}{2} \log \frac{\rho^2}{D}$, T_1 grows exponentially in n for all $a^2 < \rho^2 n^{-2/n}$. When $b > 2$, $T_2 = L^{b-1}$ grows faster than $n = bL \log L/R$. For

$$(b - b_{\min}) \left(1 - \frac{1}{R}(1 - D/a^2)\right) > 1,$$

T_3 also grows faster than n . This corresponds to the minimum value of b specified in the statement of the theorem. Therefore, under this condition, the probability of error for large n is dominated by the first term in (32). This completes the proof.

V. CONCLUSION

We have studied a new ensemble of codes for Gaussian source coding, constructed using the framework of sparse linear regression. The codewords are structured linear combinations of elements of a dictionary; the size of the dictionary is a low-order polynomial in the block length. We showed that this ensemble achieves the optimal error exponent with ML encoding for all distortions below $\frac{\sigma^2}{4.91}$, or equivalently for rates higher than 1.15 bits per source sample. This value may be an artifact of some looseness in our bounding techniques, especially in analyzing the λ^2/Δ term of Suen's inequality. We also expect that the minimum value of b required by the theorem can also be tightened by using a tighter large deviation bound for Δ . The final goal is to develop computationally feasible encoding algorithms that rapidly approach the rate-distortion bound with growing block length.

REFERENCES

- [1] A. Barron and A. Joseph, "Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity," *To Appear in IEEE Trans. Inf. Theory*. Also in Proc. 2010 IEEE ISIT.
- [2] A. Barron and A. Joseph, "Toward fast reliable communication at rates near capacity with Gaussian noise," in *2010 IEEE ISIT*. Also Yale Dept. of Stat. Technical Report, 2011.
- [3] I. Kontoyiannis, K. Rad, and S. Gitzenis, "Sparse superposition codes for gaussian vector quantization," in *2010 IEEE Inf. Theory Workshop*, p. 1, Jan. 2010.
- [4] S. Janson, *Random Graphs*. Wiley, 2000.
- [5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, Inc., 2001.
- [6] M. Eyuboglu and J. Forney, G.D., "Lattice and trellis quantization with lattice- and trellis-bounded codebooks-high-rate theory for memoryless sources," *IEEE Trans. Inf. Theory*, vol. 39, pp. 46–59, Jan 1993.
- [7] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1250–1276, June 2002.
- [8] A. Gupta, S. Verdú, and T. Weissman, "Rate-distortion in near-linear time," in *2008 IEEE Int. Symp. on Inf. Theory*, pp. 847–851.
- [9] I. Kontoyiannis and C. Gioran, "Efficient random codebooks and databases for lossy compression in near-linear time," in *IEEE Inf. Theory Workshop on Networking and Inf. Theory*, pp. 236–240, June 2009.
- [10] S. Jalali and T. Weissman, "Rate-distortion via Markov Chain Monte Carlo," in *2010 IEEE Int. Symp. on Inf. Theory*.
- [11] S. Korada and R. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Trans. Inf. Theory*, vol. 56, pp. 1751–1768, April 2010.
- [12] D. Sakrison, "A geometric treatment of the source encoding of a Gaussian random variable," *IEEE Trans. Inf. Theory*, vol. 14, pp. 481–486, May 1968.
- [13] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime: Gaussian source," in *2011 IEEE Inf. Theory Workshop*.
- [14] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 20, pp. 197–199, Mar 1974.
- [15] S. Ihara and M. Kubo, "Error exponent for coding of memoryless Gaussian sources with a fidelity criterion," *IEICE Trans. Fundamentals*, vol. E83-A, p. 18911897, Oct. 2000.
- [16] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 1998.
- [17] R. R. Bahadur and R. R. Rao, "On deviations of the sample mean," *The Annals of Mathematical Statistics*, vol. 31, no. 4, 1960.